The Serials Librarian, 58:141–148, 2010

ISSN: 0361-526X print/1541-1095 online
DOI: 10.1080/03615261003622957

Routledge
Taylor & Francis Group

From Federated Search to the Universal Search Solution

ROBERTA F. WOODS
Presenter

This presentation detailed an investigation by law librarians at Franklin Pierce Law Center Library of the viability of federated search and enterprise search to meet the needs of law library patrons. The results of the investigation revealed that federated search products could not answer the needs of law libraries to provide a single search box with a single, integrated result set because it has a fundamental flaw—it searches in real time. Neither could enterprise search solutions like the Google Search Appliance. With the backing of the New England Law Library Consortium (NELLCO) and a two-year leadership grant from the Institute for Museum and Library Services (IMLS), a committee composed of law librarians from law schools throughout New England, a vendor representative, and a development partner, IndexData, created the Universal Search Solution (USS). The USS is an open-source, standards-based, single search box solution for law libraries that may ultimately guide all libraries in similar pursuits.

KEYWORDS federated search, enterprise search, resource discovery, New England Law Library Consortium (NELLCO), Universal Search Solution (USS)

The advent of Google made one-box searching easy with result sets that seemed to be precisely what the searcher had in mind. Thus, the "Googlized" library patron was born. This patron—our patrons—will no longer tolerate anything more complex than a single search box and a single, integrated result set. In paraphrasing Herbert S. White, former dean of Library and Information Science at Indiana University, Roy Tennant wrote, "Only librarians like to *search*; everyone else likes to *find*."[1] Every day law school students, such as those at the Franklin Pierce Law Center, reinforce this concept by *not* learning how to search in the increasing variety of proprietary online databases we offer and *not* searching our catalog for books

on the shelf. Perhaps all we have done with our books is made the décor pleasing for students to study. For many law students, if it is not in Westlaw, LexisNexis, or Google, then the information does not exist. While student patrons do not necessarily say these words, their actions speak volumes.

Law students willingly learn how to use Westlaw and LexisNexis because these companies train them on how to use these services and offer free printing and prizes for learning the complexities of researching on their services. Given the smorgasbord of legal research choices offered by these two services, it is little wonder why law students believe everything legal is on Westlaw or LexisNexis and what is not there you can find on Google.

The landscape of legal information can be compared to a mall—an information mall, if you will. The mall's anchor stores are Westlaw, LexisNexis, and Google. The small stores between the anchors, the boutiques, are typically niche publishers, many with otherwise difficult to find digitized information. Boutiques have a distinct problem these days. They are typically undiscovered. Although many of these "boutique" services have very sophisticated search interfaces, they go largely unused because patrons, especially student patrons, do not have the time to learn how to effectively search them.

Simply put, the panoply of choices from which to discover digitized content requires learning how to search in numerous *un*-"Googlesque" systems and without prizes. Current library budgets cannot continue to maintain resources no one uses in the hope that one day users will seek out the content contained in them. Thus, the need for a federated system for searching across various forms of digitized content had its genesis.

FEDERATED SEARCH EXAMINED

An investigation of federated search by law librarians at Franklin Pierce Law Center was conducted prior to the creation of the New England Law Library Consortium's (NELLCO's) Universal Search Solution (USS).[2] They found that the problem with federated search in its current incarnation and various forms created by integrated library system vendors is one that cannot be overcome merely by using a different style of facets or a prettier user interface. It is the paradigm on which federated search was built, that of real-time searching, which is the problem.

Federated search grew out of the National Information Standards Organization's Metasearch Initiative to provide metasearch services, the Z39.50 standard. Originally this standard was intended to let one library system's users search the contents of other libraries with different systems. However, information providers have not adhered to this standard, resulting in systems that require proprietary search engines to access content stored in online databases.[3]

Federated search begins with a query that is sent to various databases in real time. Result sets are returned and compiled as the user waits and watches. Even after having in-house demonstrations by nearly all of the federated search vendors, it remains unclear how federated search systems actually search in the various databases. Are the queries keyword searches or are they subject or title searches? And, are some databases preferenced over others? If an information provider offered the federated search vendor incentives to allow its content to rise to the top of the list would the library be able to detect this preference and overcome it? Presumably that information would not be shared with the library by the federated search vendor since the vendor would have a profit motive to keep such information confidential. What about duplicate results? No one wants a page of the same item. Federated search cannot accomplish de-duplication of results easily or accurately because the results are gathered and displayed in real time. Latency and connectivity become issues that relate to relevancy and ranking of results. Which result is on top? Is it the first in time, or the most relevant?

The cost of a federated search system is significant for a library and desirable features, such as statistical packages, may come only with additional expense. With regard to usage statistics there is a fundamental problem with individual vendor statistics once a federated search is in use. Each and every search is logged. It is impossible to determine if the statistics include any actual retrievals of information because each and every search for a result set is counted. Thus, usage statistics are skewed. With federated search queries skewing actual usage statistics, libraries are offered no better information on product usage than before the federated search system was installed.

Displaying federated search results on the screen in real time causes the results to shift and reorganize as results are returned. The jerky display can be disconcerting to the researcher who sees something of interest flash by quickly and then is unable to locate the item of interest when the display stops moving. If no results are found in a particular database is it because there really were not any results or was the server down momentarily?

These issues of relevancy ranking, de-duplication, cost, statistics, and issues involving latency and connectivity remain unresolved, and the author believes they cannot be overcome by federated search as it is implemented today. So, federated search systems have been unable to meet the challenge for libraries and researchers, but what about enterprise search systems? Are they any better?

ENTERPRISE SEARCH EXAMINED

After investigating federated search and deciding it could not be sustained because of its flawed paradigm, enterprise search came to the forefront.

Made by a few companies including Google, an enterprise search appliance is a knowledge management tool used by large, mostly Fortune 500 companies to discover content hidden in databases and on hard drives throughout an organization. Google provided their Google Search Appliance (GSA) for testing. On first glance the GSA seemed to be the answer to our search dilemma because instead of searching in real time, the GSA uses the famous Google algorithms to create a searchable index of content. There are several methods one can use to index content: an XML feed, use the Google spider (preferred), or gather it from a database.[4]

At Franklin Pierce Law Center, we trialed the GSA and found that MAchine-Readable Cataloging (MARC) records could be converted to an XML format and fed into the appliance. Free MARC–XML conversion tools are available on the Internet. We convinced three legal publishing vendors, HeinOnline, Law Library Microform Consortia, and Oceana (now part of Oxford University Press), to go along with us on the trial and they each put up sample content on their Web servers that could be crawled. Free websites were crawled and added to the index. Administering the GSA was not difficult. The search was fast and it was possible to customize the sections of the page where sponsored results are displayed in a regular Google search. However, there were problems with the results. Since Google uses a page ranking algorithm to determine relevancy, our catalog records were always at the bottom of the list since they did not link to any other pages.

There was also the matter of price. The GSA is priced based on how many "documents" it can store rather than on storage capacity. But, what was a document? A catalog record is small compared to a full-text e-book or a journal. Was each catalog record a document? If so, could the GSA's storage capacity be maximized if the whole catalog could be defined as one document? Document storage capacity is closely tied to price so the ability to define variable sizes for documents based on the type of document would defeat Google's pricing structure, and Google was unwilling to change its marketing model for us. Given the exponential increase with which our vendor partners were adding digitized content, we would not be able to financially sustain the GSA. Although we would own the device, the software had a two-year license that would have to be renewed. Thus, the GSA answered our question about how best to create a federated search system (that is, to use an index), but as the index grew, it would quickly reach an unsustainable cost. And, the issue of our catalog records languishing at the bottom of the search results was an unacceptable result that remained unsolved.

NELLCO'S UNIVERSAL SEARCH SOLUTION

Following the investigation into both federated search and enterprise search systems, we concluded that the goal of offering our patrons a one-box

search solution to search all of the library's content could be realized only by building it ourselves. It was not a project for an independent academic law library like Franklin Pierce Law Center Library to undertake. It was more suited to a consortium so that each library did not have to reinvent and re-index the same content. The New England Law Library Consortium has twenty-five full members, including most of the academic and state law libraries located in New England. There are sixty-five affiliate members from thirty-three states across the United States, and nine international affiliate members. With support and leadership provided by Tracy Thompson-Pryzlucki, executive director of NELLCO, the Institute of Museum and Library Services (IMLS) awarded NELLCO a leadership grant in December 2007 to create the Universal Search Solution. A committee was formed to set the specifications for the USS. It consisted of law librarians representing large academic law libraries, small academic law libraries, and public law libraries throughout New England; the software developer, IndexData; and a vendor partner, HeinOnline.[5] In late March 2009, the USS was rolled out to twenty-seven law libraries who agreed to participate in the beta testing of the USS.

THE UNIVERSAL SEARCH SOLUTION EXAMINED

The difference between a search in a federated system and the USS is analogous to taking a book in hand and searching through its pages (federated search) versus searching the book's index (USS). Obviously, searching the pages is much more cumbersome and it would be easy to miss a term on a particular page. An index gives more relevant results with search terms listed once, followed by the page numbers where each term appears. A comprehensive search can be achieved by looking in one place, the index. Thus, the USS is a system that can easily de-duplicate search results since they exist in one place already organized.

NELLCO's USS is an index of content from library catalogs, proprietary databases, librarian-vetted free websites, and locally developed content. It is an open-source solution. Open source is all the rage today, and with good reason. It allows libraries to collaborate and share system development and that, in turn, leads to cost savings and a better-tailored product. The USS is being developed as an open-source, standards-based solution. Using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) standard gives content providers a protocol for their data to enable the USS to add it to the index.[6] The MARC standard allows libraries to upload a MARC file to a Web server where it is harvested by the USS on a pre-arranged schedule. Harvesting schedules are flexible and can be configured to conform to a dynamic, frequently changing catalog by harvesting the MARC records file either daily, weekly, or monthly. Librarian-vetted free websites are provided by Washburn University Law Library and crawled by the USS.

The USS is controlled by administration modules: a global administrative module used only by NELLCO and a library administrative module for individual libraries. In the global or NELLCO administrative module, all of the resources providing content to the index are listed, all of the libraries are also listed with links to their administrative modules, and the harvester is available to add resources to the index. Individual library administrative modules include a listing of all of the resources indexed in the harvester and a library selects those for which it has access by simply checking the box next to the resource. Libraries can also rename a resource for their patrons by editing the resource and supplying an alternate name for it. At Franklin Pierce Law Center we call our online catalog MelCat and prefer the search results to show "MelCat" rather than the longer name of "Franklin Pierce Law Library Catalog." This resource name change affects only our patrons' result set. In the individual library administrative modules, libraries also supply IP ranges, a logo URL, and log-in credentials for both the search interface and the administrative module.

Libraries can have more than one USS account. Having two accounts allows any participating library to set up a default search for library patrons, and a search solution for librarians that includes other local law libraries. The default search typically includes only the resources available to a researcher at their base or home institution such as the online public access catalog, proprietary content, and free websites. The librarians' search typically would include other library catalogs in the result set to aid in collection development and collaborative, regional development.

The search interface offers "Googlesque" results with the user's search terms displayed in bold in the snippet. Duplicate records are shown in the links that follow the snippet. On the right-hand side of the page source, subject and author facets are shown. Selecting a subject facet, for example, is equivalent to adding a Boolean "and" to the search with the subject facet following as the term. Drilling down further in the facets adds more Boolean "ands" to the search so it is a passive way to create a Boolean search for researchers who may not know how to construct one. The simple search box also allows researchers who know Boolean logic to construct a query in the search box. The advanced search page is linked to the simple search page and offers additional fields for searching and a way of limiting results to a particular type of content, such as articles and databases, library catalogs, vetted free websites, or scholarly repositories. The USS is accessible from NELLCO's website.[7]

Relevancy ranking of USS results is determined by the TFIDF (term frequency-inverse document frequency) algorithm. This algorithm is a standard algorithm that determines relevance based on a statistical model of counting the number of times the search term is found in a document and throughout the index. Mathematically it is expressed in its simplest form as TFIDF = (C / T) * (D / DF). As its term implies, term frequency (TF) is

essentially a percentage. It is mathematically expressed as C / T, where C is the number of times a term appears in a document and T is the total number of terms in the same document. Inverse document frequency (IDF) accounts for the fact that many words occur many times in many documents. Expressed mathematically as D / DF, where D is the total number of documents in a corpus (e.g., a database, an index, or a subset) and DF is the number of documents in which the term is found. The two quotients are multiplied together to arrive at TFIDF. A high TFIDF implies a strong relationship and documents with high TFIDF values rank higher on the list than results with lower values.[8]

Authentication for proprietary or subscribed content takes place on two levels. Initially the patron is authenticated by the USS. This happens seamlessly via IP recognition or by logging in with the appropriate credentials. When a patron clicks on a link to proprietary content, he or she is authenticated by the vendor site in the manner already established between the library and the vendor site, whether through IP recognition or username and password. A click-through to retrieve content then gets logged on the vendor site and vendor-supplied statistics are reflections of actual use of the content rather than mere searches.

The IMLS grant period runs through November 30, 2009. NELLCO continues to add resources to the index and IndexData tweaks the system based on data returned by the beta test libraries. The most glaring unresolved issue is the inclusion of providing proxy re-write capability. Without it the USS will likely be less used since our students will not be able to access content off campus or even from our wireless network in the building.

CONCLUSION

While federated search systems held the promise of a single search box and an integrated result list, they failed when it came to performance. Enterprise search offered a better search box, but was not financially sustainable and the algorithm would need to be tweaked to allow catalog records to better integrate into the search results. Creating the USS will eventually benefit all libraries because it is an open-source solution based on freely available standards.

NOTES

1. Roy Tennant, "Avoiding Unintended Consequences," *Library Journal* 126, no. 1 (2001): 38.

2. The "One Box Team" at Franklin Pierce Law Center Library consisted of Melanie Cornell, Kathy Fletcher, Barry Shanks, and Roberta Woods, who served as team leader. This team was created by Law Library director, Judy Gire, to investigate federated search and determine which product would best meet the library's needs.

3. Lynda Moulton, "Federated Search: Setting User Expectations," *Enterprise Search Blog*, February 24, 2009, http://gilbane.com/search_blog/2009/02/federated_search_setting_user_expectations.html (accessed June 9, 2009).

4. Google, "Google Search Appliance," http://www.google.com/enterprise/search/gsa.html (accessed June 8, 2009).

5. Originally the USS Committee consisted of Tracy Thompson-Pryzlucki (NELLCO executive director); Cindy Adams (NELLCO); Roberta Woods, Chair (Franklin Pierce Law Center); Leslie Rich (New York University School of Law Library); Lucinda Harrison-Cox (Roger Williams University School of Law Library); Cynthia Lewis (Vermont Law School); Simon Canick (University of Connecticut School of Law); and Martha Santoro (Massachusetts Trial Court Libraries). Both Canick and Santoro dropped out of the project due to job changes and they were replaced by Deena Frazier (Boston College) and Denise Jernigan (Connecticut State Library). Also, Cynthia Lewis alternated with Lisa Donadio of Vermont Law School when her teaching schedule conflicted with meetings. The USS Committee also includes Sebastian Hammer and David Dorman from IndexData, and Dan Rosati from HeinOnline, a valued vendor partner. For much of the development time graduate students Matthew and Kim Graff from Syracuse University School of Information Studies interned with the committee. Their contribution to the project was invaluable.

6. Open Archives Initiative, "Open Archives Initiative Protocol for Metadata Harvesting," http://www.openarchives.org/pmh/ (accessed June 8, 2009).

7. NELLCO, "Universal Search Solution," http://uss1.indexdata.com/ (accessed January 4, 2010). To log in, enter the username NELLCO and the password nellco.

8. For a good explanation of this algorithm see: Eric Lease Morgan, "TFIDF In Libraries: Part I of III (For Librarians)," *Infomotions Mini-Musings*, April 13, 2009, http://infomotions.com/blog/2009/04/tfidf-in-libraries-part-i-for-librarians/ (accessed June 9, 2009).

CONTRIBUTOR NOTE

Roberta F. Woods, J.D., served as an Assistant Professor and Reference & Electronic Services Librarian at Franklin Pierce Law Center, an independent law school located in Concord, New Hampshire, at the time of this presentation. Professor Woods also chaired the NELLCO Universal Search Solution Committee. She is now a Reference and Instructional Services Librarian at the William S. Richardson School of Law, University of Hawaii, Mānoa.